基于支持向量机的股价反转点预测①

黄朋朋　韩伟力　(复旦大学 软件学院 上海 200433)

摘　要： 股票市场瞬息万变，股价反转点对投资者进行投资决策起着至关重要的作用。技术分析能够揭示股价反转的某些特征，但是使用单一技术指标预测股价反转点的召回率和准确率不高。本文提出了一种使用支持向量机（SVM）对技术指标组合进行数据挖掘，从而实现预测股价反转点的方法，实验结果表明该方法较使用单一技术指标进行反转点预测在召回率和准确率方面都有极大的提高。

关键词： 支持向量机；股价反转点；股票市场

Stock Price Reversal Point Prediction Based on Support Vector Machines

HUANG Peng-Peng, HAN Wei-Li (Software School, Fudan University, Shanghai 200433, China)

Abstract: The stock market changes rapidly. Thus, the stock price reversal points play a vital role in investment decisions. Technical analysis can reveal some features of stock price reversal, but to use a sole technical indicator to predict the reversal points can end with sesults that have very low recall and precision rates. In order to improve recall and precision rates, a novel method of using Support Vector Machines (SVM) is proposed to data mine a combination of technical indicators. The experiment results show that this method has a higher recall and precision rate than the original ones.

Keywords: support vector machines; stock price reversal point; stock market

1　引言

股票市场瞬息万变，投资者都希望能够预测股市未来的行情走势获得更多的收益。而为了达到这个目的，寻找股价反转点就成为了一个关键点，如果能够较早判断股价的反转点，做到高抛低吸，投资者就能够获得丰厚的投资回报。

目前较为常用的三种反转点预测分析方法分别是：形态分析、趋势分析和技术分析。形态分析的基本原理是对于股价走势图形的挖掘，发现股价走势图若呈现某种形态时，预告未来价格走势的发展。使用形态分析方法判断反转点，其优点是简单易懂，但是存在判断时容易加入个人主观判断，准确率低等缺点。趋势分析主要包括道氏理论和波浪理论。趋势分析认为，价格走势存在着某种趋势，并且在没有证据来确认趋势改变之前，应假定趋势未改变；趋势周而复始，重复发生。趋势分析也存在着因为主观判断的差异而导致判断结果迥异的情况。技术分析是通过总结股市历史变化规律，把股市的各种变化数值化，使其容易被投资者接受。使用技术指标判断反转点时，通常会根据技术指标数值判断当前状态是否为超买或超卖，以此判断股价是否已经到了反转临界点。较常用的判断反转点的技术指标包括平滑异同平均线（Moving Average Convergence Divergence，MACD）、随机指标（KDJ）、相对强弱指标（Relative Strength Index，RSI）等等。由于使用单一技术指标对股价反转点进行预测存在较大的误差，所以使用多个技术指标组合进行相互验证就显得特别必要。本文希望通过对这些技术组合进行数据挖掘，获得一个更佳的反转点预测模型。

支持向量机（Support Vector Machines，SVM）是一类基于统计学习的新型机器学习方法[1]。由于它采用了结构风险最小化原则，能够较好地解决小样本、

① 收稿时间:2010-01-04;收到修改稿时间:2010-03-01

非线性和高维数问题，因而具有较好的泛化能力。此前使用SVM研究股市多数集中在对未来股价的时间序列数值预测方面[2, 3]。本文将从另外一个角度对股市进行研究，通过构造一个包含多个技术指标组合的反转点判断向量，并使用 SVM 对技术指标组合向量进行数据挖掘，得到更加准确的股价反转点预测模型。

2 系统架构

图 1 是使用 SVM 进行反转点预测的系统架构图。在进行反转点预测前，首先需要获取某支股票的历史交易数据，这些数据包括每天的开盘价、收盘价、最高价、最低价等。然后对这些数据进行预处理，并在预处理后的数据上进行反转点定义，最后抽取符合要求的技术指标组合向量，并使用 SVM 进行训练和预测。各部分的具体流程和方法如下：

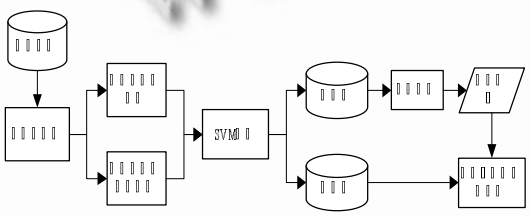


图 1　系统架构图

2.1 数据预处理

股票价格历史数据只包含了股票交易的最基本信息，要在这些历史数据上进行数据挖掘从而揭示股价反转的特征，首先需要对历史交易数据进行预处理。系统中数据预处理主要包括两方面：技术指标的计算和股票收盘价的平滑。

本系统采用了三个趋势技术指标，分别为 MACD、KDJ 和 RSI。

MACD 的计算公式如下：

$$\begin{cases} DIF_x = EMA(C_x, S) - EMA(C_x, L); \\ DEA_x = EMA(DIF_x, N); \\ MACD_x = (DIF_x - DEA_x) * 2; \end{cases} \quad (1)$$

其中 $EMA(J,T)$ 表示数值 J 的 T 日指数移动平均值，C_x 表示第 x 天的收盘价，S 表示快速指数移动平均线的天数，L 表示慢速指数移动平均线的天数，通常 S 取值为 12，L 取值为 26，N 取值为 9；

KDJ 的计算公式如下：

$$\begin{cases} RSV_x = \dfrac{C_x - RH_x}{RH_x - RL_x} * 100; \\ K_x = \dfrac{1}{3} RSV_x + \dfrac{2}{3} K_{x-1}; \\ D_x = \dfrac{2}{3} D_{x-1} + \dfrac{1}{3} K_x; \\ J_x = 3K_x - 2D_x; \end{cases} \quad (2)$$

其中 C_x 表示第 x 天的收盘价，RH_x 表示 x 天内的最高价，RL_x 表示 x 天内的最低价，通常 x 取值为 9。

RSI 的计算公式如下：

$$\begin{cases} RS_x = \dfrac{MEANUP(C,x)}{MEANDOWN(C,x)}; \\ RSI_x = 100 - \dfrac{100}{1 + RS_x}; \end{cases} \quad (3)$$

其中 $MEANUP(C,x)$ 表示 x 天内上涨收盘价的平均值，$MEANDOWN(C,x)$ 表示 x 天内下跌收盘价的平均值，通常 x 取值为 12。

股票每日收盘价连线最能表征股价走势，但是由于股票市场的特殊性，股票收盘价会在趋势方向上存在着一定的波动，为了消除收盘价波动对反转点定义的影响，本文对股票收盘价进行平滑，收盘价平滑公式如下：

$$y'(n) = \frac{\sum_{i}^{1}{}_{-1} y(n+i)}{3} \quad (4)$$

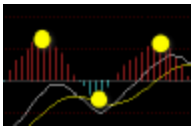
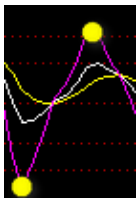
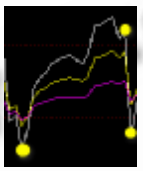
其中，$y(n)$ 是指第 n 天股票的收盘价，$y'(n)$ 是平滑后第 n 天股票的收盘价。同时，如果收盘价在趋势方向上的波动范围小于某一个足够小的给定值 $r'(r'>0)$，则忽略该小波动，具体处理方式为：假设在时间序列上有四个点 t_1、t_2、t_3、t_4，它们对应的股价分别是 Y_{t1}、Y_{t2}、Y_{t3}、Y_{t4}，则其变化率分别是：$r_1 = (Y_{t2} - Y_{t1})/Y_{t1}$、$r_2 = (Y_{t3} - Y_{t2})/Y_{t2}$、$r_3 = (Y_{t4} - Y_{t3})/Y_{t3}$，如果 $r_1 * r_3 > 0$, $r_1 * r_2 < 0$ 且 $abs(r_2) < r'$，则将变化率 r_2 置为 0。

2.2 反转点定义

本文提出的系统中存在着两类反转点：技术指标反转点和股价真实反转点。技术指标反转点通过对技术指标(包括 MACD、KDJ、RSI)的解读定义反转点，而股价真实反转点则直接从经过平滑的股票历史收盘价得到。这两类反转点的具体定义如下：

2.2.1 技术指标反转点定义[4]

表 1　技术指标反转点定义

指标	反转点定义	图标
MACD	当 MACD 柱不再创出新高，或者不再创出新低，则表明这一天是一个反转点。右图黄点标识了反转点。	
KDJ	当 $75 \le J < 100$，且不再创出新高，则标志着熊背离，是一个向下反转点；当 $0 < J \le 25$，且不再创出新低，则标志着牛背离，是一个向上反转点。右图黄点标识了反转点。	
RSI	当 $80 \le RSI < 100$，且不再创出新高，则标志着熊背离，是一个向下反转点；当 $0 < RSI \le 20$，且不再创出新低，则标志着牛背离，是一个向上反转点。右图黄点标识了反转点。	

2.2.2 真实反转点定义

对于经过平滑处理的股票收盘价序列，假设在时间序列上有三个点 t_1、t_2、t_3，它们对应的股价分别是 Yt_1、Yt_2、Yt_3，若 $(Yt_1-Yt_2)/Yt_2>R$ 且 $(Yt_3-Yt_2)/Yt_2>R$，则称 t_2 点为反转向上的反转点；反之，若 $(Yt_2-Yt_1)/Yt_2>R$ 且 $(Yt_2-Yt_3)/Yt_2>R$，则称 t_2 点为反转向下的反转点。

3　使用SVM进行反转点预测

在对股价真实反转点定义和技术指标反转点定义以后，抽取出真实反转点和技术指标反转点判断向量，使用 SVM 进行训练和测试。使用 SVM 进行训练和测试前的准备工作主要包括选择合适的核函数和得到适合样本的最优核函数参数。

3.1 SVM 原理[1]

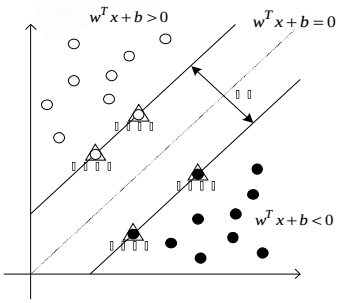


图 2　最优超平面分割

对于线性问题，如图 2 所示，在特征空间中散落着样本大小为 l 的训练样本集 $(x_1,y_1),...,(x_l,y_l), x\in R^n, y\in\{1,-1\}$，假设存在一个最优超平面将特征空间分为两个区域，对于不同类型的样本集，有如下特征：若样本 x_i 在超平面的正方，则 $y_i=1$；若样本 x_i 在超平面的负方，则 $y_i=-1$。

若超平面为：

$$w^T x+b=0 \tag{5}$$

正负样本集为：

$$\begin{cases} w^T x_i+b\ge 1, y_i=1 \\ w^T x_i+b\le -1, y_i=-1 \end{cases} \tag{6}$$

则对于每一样本，其距离分类超平面的距离为 $r=(w^T x+b)/\|w\|$，而对于支持向量来说，式（6）中的不等号应为等号，则分类间隔为：$r=\frac{1}{\|w\|}-\frac{-1}{\|w\|}=\frac{2}{\|w\|}$。

由于最佳分类平面的 r 最大，则问题可转化为：寻找 w 和使得 $r=\frac{2}{\|w\|}$ 最大，且对所有样本 $\{(x_i,y_i)\}$ 有：

$$\begin{cases} w^T x_i+b\ge 1, y_i=1 \\ w^T x_i+b\le -1, y_i=-1 \end{cases} \tag{7}$$

该式也可表达为：

$$\begin{cases} Min(\frac{\|w\|^2}{2}) \\ St: y_i(w^T x_i+b)\ge 1 \end{cases} \tag{8}$$

另外，考虑到可能存在噪音样本，这些样本会出现分类错误的问题，因此引入松弛标量 $x_i, i=1,2,...,n$ 予以解决，新的超平面的最优化问题为：

$$\begin{cases} Min(\frac{\|w\|^2}{2}+C\sum x_i) \\ St: y_i(w^T x_i+b)\ge 1-x_i, x_i\ge 0 \end{cases} \tag{9}$$

其中 $\frac{\|w\|^2}{2}$ 使样本到超平面的距离尽可能大，从而提高泛化能力；$C\sum x_i$ 则使误差尽量小，参数 C 用来调节正则化部分（$\frac{\|w\|^2}{2}$）和经验风险部分（$\sum x_i$）之间的平衡，此参数还可以看作是对错误分类点的惩罚参数，C 越大表示对错误分类的惩罚越大。

利用 Lagrange 优化方法可以把上述最优分类问题转化为其对偶问题，即：在约束条件

$$\sum_{i=1}^{n} a_i y_i=0, a_i\ge 0, i=1,...,n \tag{10}$$

下对 a_i 求解下列函数的最大值：

$$Q(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}a_ia_jy_iy_j(x_i\ x_j \qquad (11)$$

a_i 为原问题中与每个约束条件对应的 Lagrange 乘子。这是一个不等式约束下的二次函数最优化问题，存在唯一解。解中只有小部分 a_i 不为零，其对应的样本就是图 2 中所标识的支持向量。

求解上述问题后得到最优分类函数：

$$f(x)=sign\{(w^Tx+b)\}=sign\{\sum_{i=1}^{n}a_i^*y_i(x_i\cdot x)+b^*\} \qquad (12)$$

式中的求和实际上只对支持向量进行，是分类阈值，可以用任意一个支持向量且满足 $y_i(w^T\cdot x_i+b)\geq 1$ 求得。

对于非线性问题，SVM 通过非线性变换将输入空间影射到高维特征空间，然后在新空间中求解最优分类面。在最优面中采用适当的核函数 $K(x_i,y_i)$ 就可以实现某一非线性变换后的线性分类，同时没有增加计算复杂度。此时的分类函数变为：

$$f(x)=sign\{\sum_{i=1}^{n}a_i^*y_iK(x_i,x)+b^*\} \qquad (13)$$

其中的核函数需满足 Mercer 条件。

3.2 核函数选择

在支持向量机中，需要选择核函数 $K(\cdot,\cdot)$，也即一个映射 $\Phi(g)$，把 x 所在的输入空间 X 映射到另外一个空间 H，该映射只要是满足 Mercer 条件的正定函数即可。核函数主要有以下三种类型：

① 多项式核：$K(x,y)=x\cdot y$;

② RBF 核：$K(x,y)=exp(-g\left\|x-y\right\|^2)$;

③ Fourier 核：$K(x,y)=\frac{1-q^2}{2(1-2q\cos(x-y))+q^2}$

由于 RBF 核函数具有良好的性态，在实际应用中表现出了良好的性能，所以本文使用 RBF 核函数。

3.3 交叉验证选择核函数参数[5]

本文使用的 RBF 核函数有两个参数：C 和 g，本文使用 v-折交叉验证的方法，对训练集上的数据进行训练，从而找到最佳的参数对。

v-折交叉验证的思想是：把 n 个训练样本随机地分成 v 个互补相交的子集，即 $S_1,S_2,...S_v$，每折的大小相同，然后进行 v 次训练预测试，即对 S_i 进行 v 次迭代，第 i 次迭代的具体做法是：选择 S_i 为测试集，其余 v-1 个集合的并为训练集，对这两个部分的样本集进行训练和测试，得到预测准确率。经过 v 次迭代，就能够得到 v 次迭代的平均准确率。

而对于不同的 (C,g) 参数对，本文使用网格搜索的方法寻找具有最佳交叉验证平均正确率的参数对。在进行网格搜索时，可选的 C 和 g 参数值是一个指数倍增长的序列，例如：$C=2^{-5},2^{-3},...2^{15}$，$g=2^{-15},2^{-13},...,2^3$。使用每个 (C,g) 参数对在训练集上进行交叉验证后，可以得到一个最高平均准确率的参数对 (C',g')，再使用这个参数对在测试集上进行测试。

4 实验与结论

本文的实验数据来源于 Wind 咨询金融终端[6]，并以上证指数为实验对象，选择 1999 年 1 月 27 日至 2009 年 10 月 14 日共 2581 个交易日的历史数据进行实验。对这些数据进行预处理和反转点定义，对于第 i 天的交易数据，可以得到如下向量：

$$x_i:x_{i1},x_{i2},x_{i3} \qquad (14)$$

x_i,x_{i1},x_{i2},x_{i3} 分别表示该天是否真实反转点、是否被 MACD 指标判断为反转点、是否被 KDJ 指标判断为反转点、是否被 RSI 指标判断为反转点。如果是反转点，则对应向量值为 1，否则为 0。

如果某一天的向量为 $x_j:x_{j1},x_{j2},x_{j3}$，若 $x_{j1}+x_{j2}+x_{j3}>0$，则至少有一个技术指标表示该天为反转点，将所有具有该性质的向量抽取出来以供 SVM 训练和预测，共提取出 284 条向量。

将这 284 条向量中的 212 条作为训练集，72 条作为测试集，使用 libsvm[7]进行训练和测试。在对训练集进行交叉验证和网格搜索后，得到最佳的 (C,g) 参数对为 $(2^{-1},2^{-1})$，并使用该参数对在测试集上测试，得到 SVM 预测值。

可以很容易地算得使用单一技术指标（MACD、KDJ、RSI）以及用 SVM 对这三个技术指标组合向量进行反转点预测实验的召回率（Recall Rate）、准确率（Precision Rate）和 F-Measure，这三个衡量指标的计算公式如下：

$$\begin{cases} recall=\dfrac{a}{a+c}; \\ precision=\dfrac{a}{a+b}; \\ F-mesure=\dfrac{2*precision*recall}{precision+recall} \end{cases} \qquad (15)$$

其中 a 表示该天是真实反转点，且被指标判断为反转

点的样本天数总和；b 表示该天被指标判断为反转点，但不是真实反转点的样本天数总和；c 表示该天是真实反转点，但是没有被指标判断为反转点的样本天数总和。实验结果如表 2 所示：

表 2　实验结果

指标	Recall	Precision	F-Measure
MACD	47.91%	67.64%	56.10%
KDJ	39.58%	65.52%	49.35%
RSI	60.42%	63.04%	61.70%
SVM	83.33%	71.43%	76.92%

从实验结果可以看到，使用 SVM 进行反转点预测，较单一使用 MACD、KDJ、RSI 技术指标，召回率分别提高了 35.42%、43.75%、22.91%；准确率分别提高了 3.79%、5.91%、8.39%；F-Measure 分别提高了 20.82%、27.57%、15.22%。这说明使用 SVM 进行反转点预测较单一指标能够识别更多的反转点，并对这些反转点给出较高的判断结论，以便于投资者进行投资决策。

5　结束语

本文讨论了股价反转点预测的问题，针对使用单一技术指标预测法召回率和准确率不高的情况，本文使用支持向量机对多个技术指标组合进行数据挖掘，以期获得更高的召回率和准确率。对上证指数 10 年的实验表明，基于 SVM 的股价反转点预测具有更高的召回率和准确率。

参考文献

1 Naumovich VV. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

2 Kyoung-Jae K. Financial time series forecasting using support vector machines. Neurocomputing. 2003, 55(1-2):307－319.

3 Tay Francis EH，Cao LJ. Modified support vector machines in financial time series forecasting. Neurocomputing. 2002,48(1-4):847－861.

4 Alexander E. 以交易为生. 北京: 机械工业出版社, 2009.

5 Hsu CW，Chang CC，Lin CJ. A Practical Guide to Support Vector Classification. 2009.

6 Wind 咨询金融终端. http://www.wind.com.cn/.

7 Chang CC，Lin CJ. LIBSVM -- A Library for Support Vector Machines. 2001.